Speech Communication 112 (2019) 15–21

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A statistical procedure to adjust for time-interval mismatch in forensic voice comparison

Geoffrey Stewart Morrison[a,b,*], Finnian Kelly[c]

[a] Forensic Speech Science Laboratory, Aston Institute for Forensic Linguistics, and Centre for Forensic Data Science, Department of Computer Science, Aston University, Birmingham, United Kingdom
[b] Forensic Evaluation Ltd., Birmingham, United Kingdom
[c] Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX, United States

ARTICLE INFO

Keywords:
Forensic voice comparison
Likelihood ratio
Time interval
Mismatch

ABSTRACT

The present paper describes a statistical modeling procedure that was developed to account for the fact that, in a forensic voice comparison analysis conducted for a particular case, there was a long time interval between when the questioned- and known-speaker recordings were made (six years), but in the sample of the relevant population used for training and testing the forensic voice comparison system there was a short interval (hours to days) between when each of multiple recordings of each speaker was made. The present paper also includes results of empirical validation of the procedure. Although based on a particular case, the procedure has potential for wider application given that relatively long time intervals between the recording of questioned and known speakers are not uncommon in casework.

1. Introduction

The present paper describes a statistical modeling procedure that was developed to account for the following situation in a forensic voice comparison analysis:[1]

- There is a long time interval (e.g., years) between when the questioned-speaker recording was made and when the known-speaker recording was made.[2]
- In the sample of the relevant population used for training and testing the forensic voice comparison system there is a short interval (e.g., hours or days) between when each of multiple recordings of each speaker was made.

Although originally developed for a particular case,[3] the procedure has potential for wider application; in the first author's experience it is not uncommon to receive requests to conduct casework involving time intervals of several years between when the questioned-speaker and known-speaker recordings were made. It is impractical to collect training and test data that include time intervals of this length and that also represent the relevant population and reflect the speaking styles and other recording conditions in the case.

Our purpose in the present paper is to describe and validate the statistical modeling procedure in general, not to rework the original case. The procedure we describe in the present paper has been further developed and refined since its application in the original case, and, although similar, the forensic voice comparison system used for the research reported here is not identical to that used for the case. Also, the research reported here does not make any use of the case-specific recordings that were used for the analysis in the case. The latter were provided with the stipulation that they only be used for conducting the analysis in that case. The description immediately below of the case conditions and of the system used to conduct the analysis in the case is therefore deliberately terse.

In the original case, approximately six years had elapsed between when the recording of the questioned speaker was made and when recordings of the known speaker were made. The questioned-speaker recording was of a telephone call made to a police call center. The known speaker used multiple mobile handsets to make multiple calls to the call center over several days, and was recorded on the same equipment as had been in use six years previously (8 usable recordings were obtained). Just over 100 other speakers also used multiple mobile handsets to make multiple calls over several days to the same police call center and were recorded on the same equipment (at least 5

[*] Corresponding author.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

[1] We assume a reader familiar with the likelihood ratio framework for the evaluation of evidence and with human-supervised automatic approaches to forensic voice comparison. Readers unfamiliar with these topics may wish to consult Morrison et al. (2018) and Morrison et al. (in press).

[2] "Questioned speaker" and "known speaker" are used as abbreviations for the speaker of questioned identity and the speaker of known identity respectively.

[3] *R v Dunstan* [2018] ONSC 4153. Other aspects of the forensic voice comparison testimony in that case are discussed in Morrison and Enzinger (2019).

https://doi.org/10.1016/j.specom.2019.07.001
Received 6 March 2019; Received in revised form 24 June 2019; Accepted 2 July 2019
Available online 2 July 2019

usable recordings were obtained from each speaker). The latter speakers (hereinafter “sample speakers”) were recruited such that they were representative of the relevant population for the case (adult male speakers of General Canadian English from southern Ontario).[4] The questioned-speaker recording was short, resulting in approximately 10 s net speech from the speaker of interest. The known speaker and the sample speakers were asked to memorize and repeat a short script that contained the same phrases as had been spoken by the questioned speaker. Mismatches between the questioned-speaker recording and the known- and sample-speaker recordings were therefore minimal, except for the six year time interval between the questioned-speaker recording and the known-speaker recordings versus intervals of hours to days between the multiple recordings of each sample speaker.

The forensic voice comparison system used was similar to that described in §2.2 below. The system was an identity vector - probabilistic linear discriminant analysis (i-vector PLDA) system (Dehak et al., 2011). i-vectors from recordings of ~50 of the sample speakers were used to train a linear discriminant analysis (LDA) model and for training the PLDA model (the LDA model was used for mismatch compensation and dimension reduction). Pairs of i-vectors from recordings of the other ~50 sample speakers were passed through the LDA and PLDA models in order to generate a set of scores originating from same-speaker comparisons and a set of scores originating from different-speaker comparisons. These same-speaker and different-speaker scores were then used to train a regularized logistic regression model to convert scores to likelihood ratios (see: Pigeon et al., 2000; González-Rodríguez et al., 2007; Morrison, 2013; Morrison and Poh, 2018). The same set of scores was used for empirical validation (to avoid training and testing on the same data, cross-validation was used for the logistic regression model).

If no additional steps had been taken, and the procedure described above had been used to calculate a likelihood ratio for the comparison of the questioned- and known-speaker recordings in the case, the result would have been biased and misleading. It would have been biased and misleading because the training data did not have the same six-year time interval as existed between the questioned- and known-speaker recordings. The time intervals between the multiple recordings of each speaker in the training data ranged from only a few hours to a few days. Examples of such biased and misleading results are provided in §4 below.

Kelly and Hansen (2016) described a procedure for calibrating an automatic speaker verification system when there are differing time intervals between enrollment and verification. That paper’s Fig. 3 showed the distributions of different-speaker scores and same-speaker scores resulting from comparisons made across a range of time intervals. A visually salient pattern in that figure was that as the time interval increased the values of the same-speaker scores decreased and moved closer to the different-speaker scores (see also the example given in Fig. 3 of the present paper). This observation provides the basis for the procedure presented in the present paper for accounting for the mismatch in the time interval between the questioned- and known-speaker recordings versus the time interval between the multiple recordings of each of the speakers in the data used for training and testing the system. Each same-speaker score used for training and testing was derived from a pair of recordings made only hours to days apart. The idea is to decrease the values of those same-speaker scores so that their distribution reflects what would be expected if each same-speaker score were derived from a pair of recordings made six years apart. The procedure is similar to the within source degradation procedure described in González-Rodríguez et al. (2006), but the degree of decrease of the same-speaker scores is determined via empirical analysis of scores derived from a database of recordings that includes a range of time intervals between multiple recordings of each of multiple speakers. We call it a “time-interval-adjustment procedure”, which makes use of a “time-interval-adjustment model”.

The remainder of the present paper is organized as follows:

- Section 2 describes the time-interval-adjustment procedure and the data that were used to train the time-interval-adjustment model.
- Section 3 describes the results of empirical validation of the time-interval-adjustment procedure.
- Section 4 provides an example of applying the procedure.

The score data derived in the present study, and the Matlab scripts used for conducting the analyses on the score data, are available at https://doi.org/10.17036/researchdata.aston.ac.uk.00000405.

2. Time-interval-adjustment procedure

2.1. *Training data*

The data used to train the time-interval-adjustment model were taken from the Multisession Audio Research Project (MARP) corpus previously described in Lawson et al. (2009) and Kelly and Hansen (2015). The corpus includes recordings of 46 adult male and 27 adult female speakers of US English. The corpus includes recordings of multiple speaking styles. The present research makes use of the recordings of conversational speech from the 46 male speakers (each conversation was approximately 10 min long). The core MARP dataset contains recordings made at intervals of approximately two months over approximately a three-year time period. The ages of the speakers at the time of the first and last recording sessions were as given in Table 1. Recordings from a total of 19 recording sessions were available for analysis (data from the first and sixth of the original 21 recording sessions were not released). Additional recordings of a subset of the original MARP speakers were made approximately seven years after the last of the core MARP sessions (approximately ten years after the first session). All recordings were made using headset microphones in sound attenuated booths. The recording equipment and environment remained consistent across all sessions. Due to the length of time elapsed, however, there may have been some discrepancies in recording conditions. The audio was recorded as PCM at 48 kHz, 24 bit quantization, using an Edirol FA101 firewire soundcard. Recordings were subsequently downsampled to 8 kHz, 16 bit quantization.

Data from the 46 male speakers were used to calculate scores. In order to control for recording duration, each recording of each speaker in each session was split into sections of 60 s net speech, i.e., 60 s post voice activity detection (VAD). There were up to 3 non-overlapping sections per speaker per session.

2.2. *i-vector PLDA system*

Scores were generated using an i-vector PLDA system

Features were extracted after application of an energy-based VAD. Features were mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980), extracted using 32 ms wide hamming windows, step size 16 ms, 24 filters in the frequency range 1 Hz – 4 kHz, and 1st through 15th coefficients saved. Deltas and double deltas (Furui, 1986) were appended to the MFCC vectors. Deltas were calculated over the adjacent ± 4 MFCC vectors, and double deltas were calculated over

Table 1
Ages of speakers (in years) at time of first and last recording sessions.[5]

Age range:	20–30	31–40	41–50	51+
Num. speakers first session:	15	12	8	11
Num. speakers last session:	0	4	8	3

[4] For discussion of issues related to the selection of the relevant population in forensic voice comparison and the importance of training and testing using data that reflect the relevant population, see Morrison et al. (2016) and Morrison (2018).

Table 2
Number of same-speaker scores and number of speakers contributing to the same-speaker scores for each time interval. Time intervals (in months) are approximate.

Time interval:	2	4	6	8	10	12	14	16	18	20	22
Num. speakers:	46	42	42	40	40	40	39	38	38	37	36
Num. scores:	3634	3368	3165	2753	2820	2521	2290	2192	1837	1575	1521
Time interval:	24	26	28	30	32	34	36	38	82	86	92
Num. speakers:	36	33	32	28	28	22	21	15	12	14	13
Num. scores:	1130	920	765	564	635	454	282	132	105	120	105
Time interval:	94	96	98	102	104	110	114	116	118	120	
Num. speakers:	14	15	15	15	15	15	15	15	15	13	
Num. scores:	120	132	129	129	132	129	123	115	115	111	

the adjacent ± 2 delta vectors. Global cepstral mean subtraction (CMS; Furui, 1981) was used for feature-domain mismatch compensation.

The UBM had 1024 Gaussian components, and the T matrix extracted 400 dimensions (Dehak et al., 2011). LDA was used to reduce the number of dimensions to 200, and the results were length normalized. PLDA (Prince and Elder, 2007; see also Sizov et al., 2014) was applied with no additional dimension reduction. Training data for these models came from a diverse set of speech recordings from several thousand speakers. These recordings represented a wide range of microphone, and landline and mobile telephone conditions. No MARP data were used in training the i-vector PLDA system.

2.3. *Scores*

2.3.1. *Same-speaker scores*

Using the i-vector PLDA system, a score was generated for the comparison of each section of each speaker's recording from each MARP recording session with each section of the same speaker's recordings from each of the other MARP recording sessions. The time interval between each same-speaker pair of recordings was noted. Same-session comparisons were not made as these would have been same-recording comparisons.

Recordings from all 46 speakers were not available in all sessions. Table 2 provides the number of same-speaker scores available for each time interval and the number of speakers contributing to those scores. Time intervals, given in months, are approximate. Time intervals for which there were fewer than 100 scores, or for which fewer than 10 speakers contributed scores, were excluded from analysis and are not shown in Table 2.

2.3.2. *Different-speaker scores*

Using the i-vector PLDA system, a score was generated for the comparison of each section of each speaker's recording from the earliest available MARP session versus each section of every other speaker's recordings from the second earliest available MARP session. The interval between these different-speaker pairs of recordings was the same as the shortest interval for the same-speaker pairs of recordings, approximately 2 months. Barring a radical shift in the entire population's speech production, the time interval between the members of different-speaker pairs is not relevant (we envisage, however, an application in which the only scores available are based on a short time interval).

There were 9517 different-speaker scores from 38 speakers.

2.4. *Modeling the relationship between score means and time interval*

The 10% trimmed mean was calculated for the different-speaker scores, and for the same-speaker scores at each time interval. The 10% trimmed mean rather than the regular mean was used because kernel-density plots of the same-speaker scores revealed that they had low-value outliers.

The difference, d_t, between the different-speaker score mean, $\hat{\mu}_{\mathrm{ds}}$, and the same-speaker score mean, $\hat{\mu}_{\mathrm{ss},t}$, was calculated for each interval, t, see Eq. (1).

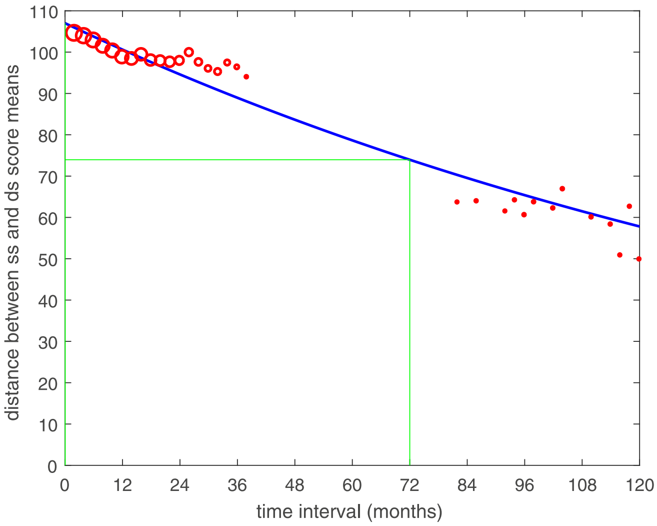

Fig. 1. Circles: Plot of distance between same-speaker score mean and different-speaker score mean, d_t, at each time interval in the training data, t. Thick line: Fitted weighted regression. The regression was weighted by the number of scores contributing to each mean, and the relative weighting is represented by the sizes of the circles. The thin vertical and horizontal lines represent the values that would be used to shift scores with a time interval of 1 day to the expected location of scores with a time interval of 6 years.

$$d_t = \hat{\mu}_{\mathrm{ss},t} - \hat{\mu}_{\mathrm{ds}} \tag{1}$$

The d_t versus t values are plotted as circles in Fig. 1. Although there is some noise, a pattern is apparent whereby as the time interval increases the distance between the same-speaker mean and different-speaker mean decreases. A weighted least-squares linear regression with an exponential link function was fitted to the d_t versus t values. Weighting was according to the number of scores for each interval (similar results were obtained if weighting by the number of speakers). The size of the circles in Fig. 1 represent their relative weights. The thick line represents the fitted regression, see Eq. (2). The fitted values for the intercept and slope coefficients were $a = 4.67$ and $b = -5.39 \times 10^{-3}$.

$$\hat{d}_t = e^{a+bt} \tag{2}$$

We also explored the relationship between the variances of the same-speaker scores and the time intervals, but concluded that there was not a systematic relationship. The time-interval adjustment model is therefore based only on the relationship between score means and time intervals. Apart from outliers (which were handled using trimmed calculations of means), and taking into account the small number of speakers contributing to scores as some intervals, plots of the score distributions appeared to be reasonably close to normally distributed (as is generally observed for the output of i-vector PLDA systems).

2.5. Adjusting same-speaker score values to account for time-interval mismatch

Our aim is to calculate the degree by which to shift same-speaker scores from pairs of recordings with time interval t_0 so that they reflect the score values that would be expected if the interval had been t_1. In Eq. (3), $\Delta_{t_0\to t_1}$ is the proportion by which to decrease the distance between the same-speaker score mean and the different-speaker score mean. Since this is a proportion, it can be applied to adjust not only the same-speaker scores generated from the MARP dataset, but also same-speaker scores generated from other datasets, such as same-speaker scores generated from case-relevant data. The formula for adjusting a score is given in Eq. (4), in which $\hat{\mu}^{\mathrm{c}}_{\mathrm{ss}}$ and $\hat{\mu}^{\mathrm{c}}_{\mathrm{ds}}$ indicate the means for same-speaker and different-speaker scores generated from case-relevant data, $x^{\mathrm{c}}_{t_0}$ is the same-speaker score value to be adjusted, and $\hat{x}^{\mathrm{c}}_{t_1}$ is the adjusted score value, i.e., the estimated value if the time interval between the pair of recordings had been t_1 rather than t_0.

$$\Delta_{t_0\to t_1} = 1 - \frac{\hat{d}_{t_1}}{\hat{d}_{t_0}} \tag{3}$$

$$\hat{x}^{\mathrm{c}}_{t_1} = x^{\mathrm{c}}_{t_0} - \Delta_{t_0\to t_1}\left(\hat{\mu}^{\mathrm{c}}_{\mathrm{ss}} - \hat{\mu}^{\mathrm{c}}_{\mathrm{ds}}\right) \tag{4}$$

The thin vertical and horizontal lines in Fig. 1 represent the t_0 value (1 day = 1/30 month) and t_1 value (6 years = 72 months) for the original case, and their corresponding $\hat{d}_{t_0} = 107$ and $\hat{d}_{t_1} = 74.0$ values. The proportional shift was $\Delta_{t_0\to t_1} = 0.309$, i.e., same-speaker scores used for training the logistic regression models and for testing the system in the case should be shifted downward in value by 30.9% of the distance between the means of the different-speaker and same-speaker scores calculated for the case. Note that the score for the comparison of the actual questioned-speaker and same-speaker recordings should not be shifted as it is actually based on a six year time interval.

3. Empirical validation

3.1. Procedure

To validate the time-interval-adjustment procedure, we conducted a series of cross-validated tests. We shifted the scores from the shortest interval represented in the MARP data to the estimated locations for each of the longer intervals. The training was leave-two-intervals out: Data from the origin time interval, t_0, and the target time interval, t_1, were excluded from training. The training was also leave-one-speaker out: For each speaker, scores from all comparisons involving that speaker were excluded from training the model used to adjust that speaker's same-speaker scores. The regression was weighted according to the number of scores remaining for each interval after the latter scores were excluded.

4. Results

In Fig. 2, the circles represent the actual distances between the different-speaker and same-speaker means at each time interval (d_t values), and the crosses represent the mean distances for the 2-month-interval data after they have been shifted to the estimated locations for each of the longer intervals ($\overline{\hat{x}^{\mathrm{c}}_{t_1}}$ values). All symbols are plotted the same size; all means were 10% trimmed. What appears to be a band is a collection of thin lines in which each line represents the fitted regression for one cross-validation run. The root-mean-square (RMS) error rate was 4.51% (expressed as a percentage of the distance between the different-speaker and the 2-month-interval same-speaker means, i.e., as a percentage of d_{t_0}). For intervals in the range 16 to 38 months, the shifts were consistently greater than necessary to exactly match the mean of the data that actually reflected those time intervals. The worst per-interval error was at the 34-month interval: 8.53% more than necessary to exactly match the mean of the data that actually reflected that interval. The error pattern was more variable in the 82-month-plus region where the data were relatively noisy due to them coming from a relatively small number of scores from a relatively small number of speakers. A more complex model could be fitted to give lower error rates, but for generalizability we considered it better to use a parsimonious model.

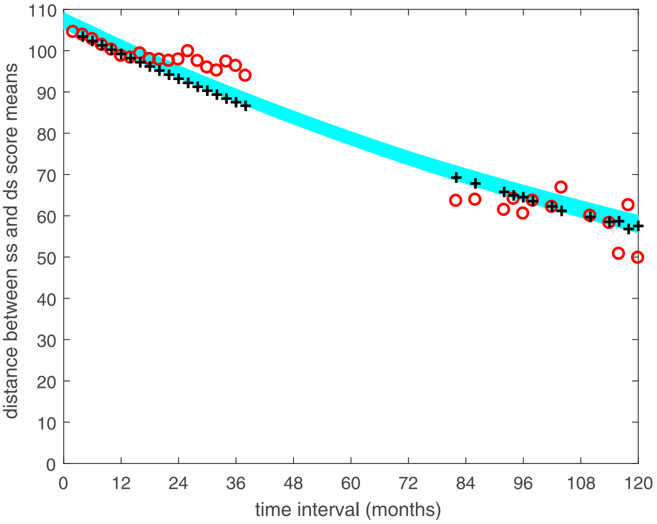

Fig. 2. Circles: Plot of distance between same-speaker score mean and different-speaker score mean, d_t, at each time interval in the training data, t. Crosses: Plot of distance between same-speaker score mean and different-speaker score mean for the 2-month-interval data after they have been shifted to the estimated location corresponding to each of the other time intervals represented in the training data. Thin lines: Fitted weighted regression for each cross-validation run.

5. Example

Fig. 3 shows an example of the adjustment of the same-speaker scores from a 2-month interval to a 102-month ($8\frac{1}{2}$-year) interval. We use this origin- and target-interval pair rather than 1-day to 6-years as in the case because we have data at the 2-month interval that we can adjust and we have data at the 102-month interval against which we can compare. The 102-month target also has one of the closest correspondences between

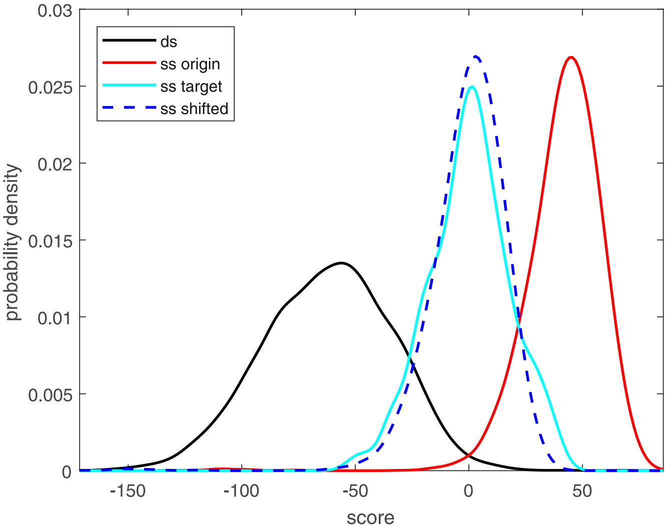

Fig. 3. Distributions of different-speaker scores (ds), 2-month-interval same-speaker scores (ss origin), 102-month-interval same-speaker scores (ss target), and 2-month-interval same-speaker scores shifted to the estimated location for the 102-month interval scores (ss shifted).

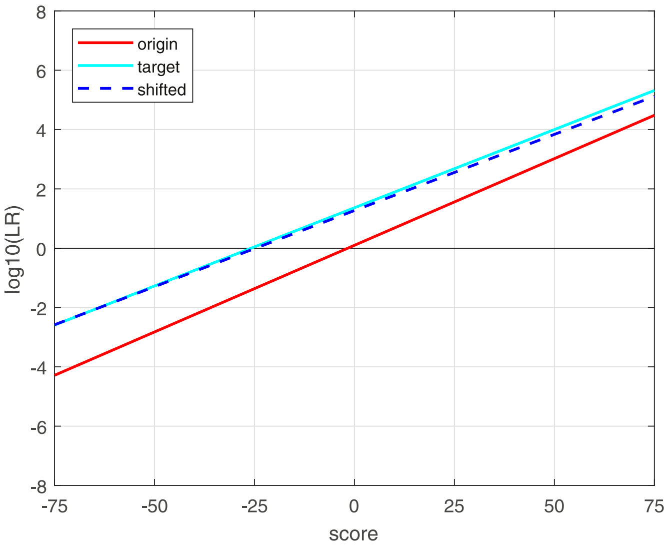

Fig. 4. Score to log likelihood ratio mappings from logistic regression models. Models trained on different-speaker scores plus: 2-month-interval same-speaker scores (origin), 102-month-interval same-speaker scores (target), and 2-month-interval same-speaker scores shifted to the estimated location for the 102-month interval scores (shifted).

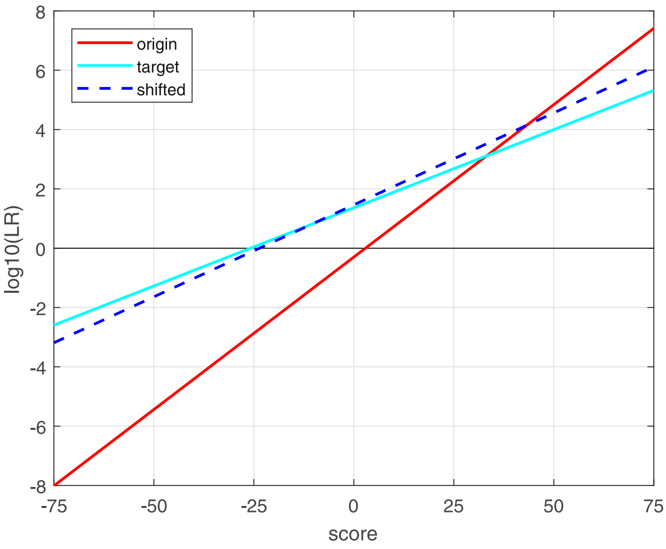

Fig. 5. Score to log likelihood ratio mappings from logistic regression models. Models trained on different-speaker scores plus: 2-month-interval same-speaker scores with the lowest 1% of scores trimmed (origin), 102-month-interval same-speaker scores (target), and 2-month-interval same-speaker scores with the lowest 1% of scores trimmed shifted to the estimated location for the 102-month interval scores (shifted).

adjusted data and data actually from the target interval – it is presented as an example not as a validation of the procedure.

For the 2-month to 102-month adjustment, $\Delta_{t_0 \rightarrow t_1}$ was 40.4%. This shift was only 0.14% less than necessary to exactly match the mean of the data that reflected an actual 102 month time interval.

Fig. 4 shows the score to likelihood ratio mapping function (calibration function) resulting from fitting a logistic regression model (without regularization) to:

- The different-speaker scores plus the 2-month-interval same-speaker scores – labelled “origin”.
- The different-speaker scores plus the 102-month-interval same-speaker scores – labelled “target”.
- The different-speaker scores plus the 2-month-interval same-speaker scores adjusted to the estimated location of a 102-month interval – labelled “shifted”.

Note that, whereas the 2-month-interval mapping function is quite far from the 102-month-interval mapping function, after time-interval adjustment is applied to the former it lies almost on top of the latter – this is the desired result.

A reduction in the separation of the different-speaker and same-speaker score sets would be expected to result in both a shift and a reduction in slope of the mapping function. The shift can be easily seen in Fig. 4, but the reduction in slope is slight because the 2-month interval mapping function already had a relatively shallow slope. This shallow slope was due to the 2-month-interval same-speaker score set having a small number of scores with very small values (about 1% of the total number of same-speaker scores at that time interval). The 2-month same-speaker data were trimmed by excluding the lowest 1% of scores. We do not recommend this for casework, but do it here for illustrative purposes (the 1% trimming of the lowest value scores here should not be confused with the 10% trimmed means used in training the time-interval adjustment model). Fig. 5 shows the score to likelihood ratio mapping function resulting from fitting a logistic regression model to the 1% trimmed data. In Fig. 5 the 2-month-interval mapping function has a steep slope. That slope is substantially reduced by the time-interval-adjustment procedure.

Fig. 6 shows Tippett plots of validation results from different combinations of training and test data. Data were not 1% trimmed. Cross-validation was used to avoid training and testing the logistic regression model on the same data: scores were excluded from the training data if one or both of the contributing speakers was the same as a speaker that contributed to the score being calibrated.

Fig. 6(a) represents the validation results that would have been obtained if the time-interval mismatch had been ignored and training and testing had been done on 2-month-interval data. The performance is extremely good (the data were high-quality audio and so not representative of casework conditions), the log likelihood-ratio cost is 0.089 (C_{llr}; Brümmer and du Preez, 2006; González-Rodríguez et al., 2007; Morrison, 2011; Meuwly et al., 2017). These results, however, would be highly misleading if there were a time-interval mismatch: If the questioned- and known-speaker recordings had a 102-month interval, then it is the results of testing with 102-month data that would be informative of performance under this condition. The results of training on 2-month-interval data and testing on 102-month data are shown in Fig. 6(b). The results are heavily biased, and the C_{llr} value is high, 0.797. The results shown in 6(a) and 6(b) empirically demonstrate that failing to take account of the time-interval mismatch would: (a) produce misleadingly good validation results; and (b) produce misleading biased likelihood-ratio values for the comparison of the questioned- and known-speaker recordings.

Fig. 6(c) represents the validation results from training and testing on 102-month data. If the questioned- and known-speaker recordings had a 102-month interval, and we had 102-month-interval data for training and testing that also reflected the relevant population and other conditions for the case, then this is what we should use for training and testing. Results are quite good (the data were high-quality audio and so not representative of casework conditions), the C_{llr} value is 0.337. Fig. 6(d) represents the validation results from training and testing on data in which the time-interval-adjustment procedure has been applied to shift the 2-month-interval same-speaker scores to the estimated location for 102-month-interval same-speaker scores. The resulting Tippett plot is very similar to that derived from actually training and testing on 102-month-interval data, and the C_{llr} value is also similar, 0.322. Some difference will be due to the fact that whereas all 46 speakers had scores at the 2-month interval, only 15 had scores at the 102-month interval. The results shown in 6(c) and 6(d) empirically demonstrate

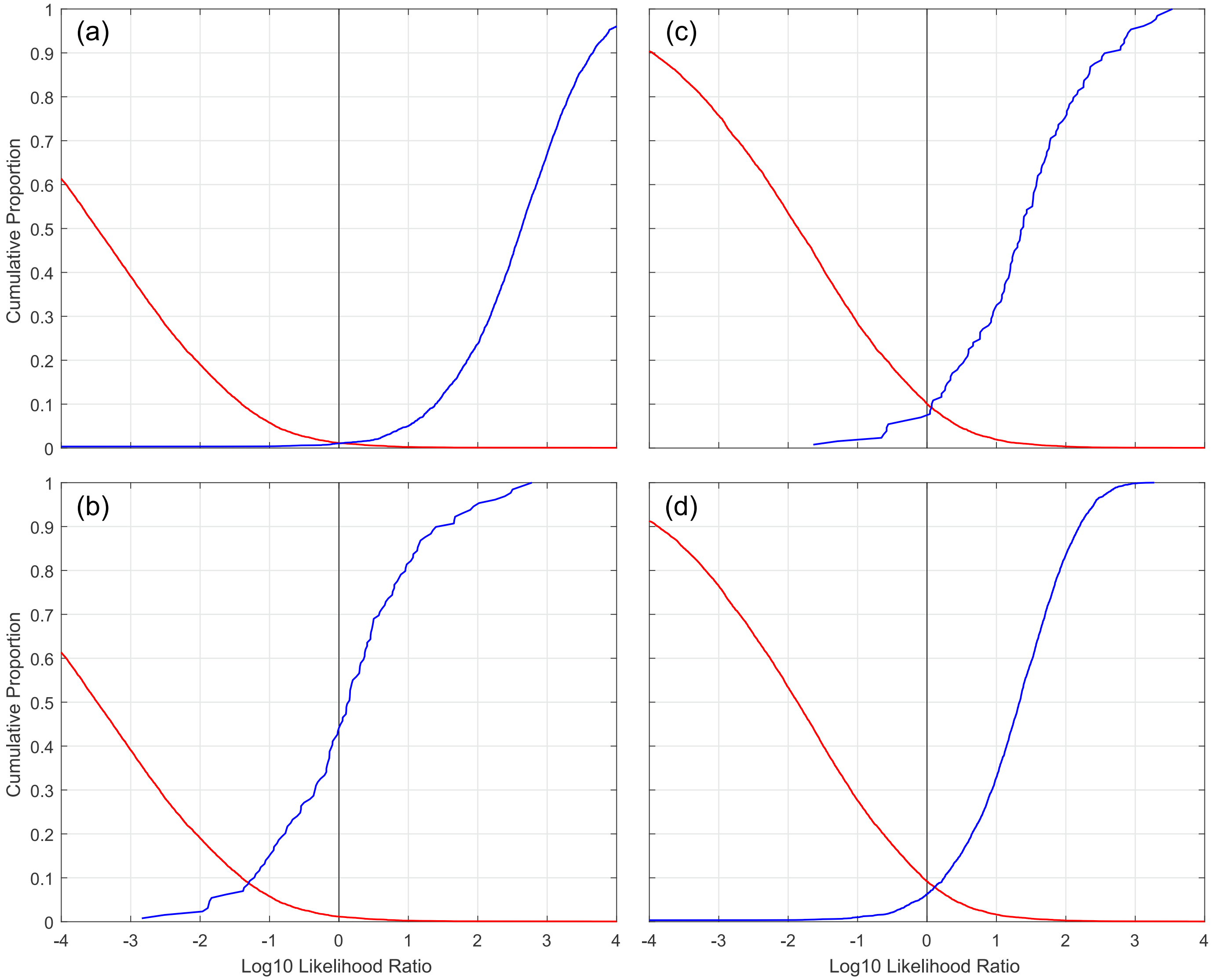

Fig. 6. Tippett plots resulting from logistic regression models: (a) trained and tested using different-speaker scores plus 2-month-interval same-speaker scores [origin]; (b) trained using different-speaker scores plus 2-month-interval same-speaker scores, but tested with different-speaker scores plus 102-month-interval same-speaker scores [mismatched]; (c) trained and tested using different-speaker scores plus 102-month-interval same-speaker scores [target]; (d) trained and tested using different-speaker scores plus 2-month-interval same-speaker scores shifted to the estimated location for the 102-month interval scores [shifted].

that the time-interval adjustment procedure (d) can be effective in replicating the effect of an actual long time interval (c) between when the questioned- and known-speaker recordings were made.

6. Discussion and conclusion

We have illustrated that if there is a large time-interval between when the questioned-speaker recording is made and when the known-speaker recording is made, but one trains and tests using pairs of same-speaker recordings that have a short time-interval between when each member of each pair was made, then:

- the validation results will be misleadingly good, and
- the likelihood-ratio value calculated for the comparison of the questioned- and known-speaker pair will be biased.

We have proposed a procedure for adjusting short-interval same-speaker scores to the values they would be expected to have if they came from a longer interval. The time-interval-adjustment procedure is based on a shift relative to the distance between the means of the same-speaker and the different-speaker scores. The idea is that because this is a relative shift (i.e., a proportion of the distance between the same- and different-speaker score means), it can be applied to datasets other than the dataset used for training the time-interval-adjustment model, and in particular it can be applied to case-relevant data. The time-interval-adjustment model was trained on a dataset consisting of recordings of multiple speakers with each speaker recorded multiple times over a period of several years. We presented validation results from cross-validation on the same dataset. We would like to perform cross-dataset validation in which, like when applied to a real case, the conditions of the test dataset differ from those of the dataset used to train the time-interval-adjustment model. Not currently having access to another suitably sized dataset with consistent short and long time intervals between recordings of each speaker, we have not yet been able to validate the cross-dataset application of the procedure. Collecting such a dataset is challenging, but we hope that this can be achieved in the future, making cross-dataset validation possible. We also note that the particular time-interval-adjustment model used in the study reported in the present paper was trained using scores generated using a particular i-vector PLDA system. Different systems for generating scores are likely to have different performance characteristics, hence the time-interval-adjustment model used for a particular case should be trained on scores generated using the same system that will be used to generate scores for that case.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Gold Open Access funding was provided by Aston University.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2019.07.001.

References

Brümmer, N., du Preez, J., 2006. Application independent evaluation of speaker detection. *Comp. Speech Lang.* 20, 230–275. https://doi.org/10.1016/j.csl.2005.08.001.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust.* 28, 357–366. https://doi.org/10.1109/TASSP.1980.1163420.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* 19, 788–798. https://doi.org/10.1109/TASL.2010.2064307.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust.* 29, 254–272. https://doi.org/10.1109/TASSP.1981.1163530.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust.* 34, 52–59. https://doi.org/10.1109/TASSP.1986.1164788.

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., Ortega-García, J., 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Comp. Speech Lang.* 20, 331–355. http://dx.doi.org/10.1016/j.csl.2005.08.005.

González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D.T., Ortega-García, J., 2007. Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Trans. Audio Speech Lang. Process.* 15, 2104–2115. https://doi.org/10.1109/TASL.2007.902747.

Kelly, F., Hansen, J.H.L., 2015. Evaluation and calibration of short-term aging effects in speaker verification. In: Proceedings of Interspeech 2015, pp. 224–228. http://www.isca-speech.org/archive/interspeech_2015/i15_0224.html.

Kelly, F., Hansen, J.H.L., 2016. Score-aging calibration for speaker verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* 24, 2414–2424. http://dx.doi.org/10.1109/TASLP.2016.2602542.

Lawson, A.D., Stauffer, A.R., Cupples, E.J., Wenndt, S.J., Bray, W.P, Grieco, J.J., 2009. The multi-session audio research project (MARP) corpus: goals, design and initial findings. In: Proceedings of Interspeech 2009, pp. 1811–1814. http://www.isca-speech.org/archive/interspeech_2009/i09_1811.html.

Meuwly, D., Ramos, D., Haraksim, R., 2017. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forens. Sci. Int.* 276, 142–153. http://dx.doi.org/10.1016/j.forsciint.2016.03.048.

Morrison, G.S., 2011. Measuring the validity and reliability of forensic likelihood-ratio systems. *Sci. Justice* 51, 91–98. http://dx.doi.org/10.1016/j.scijus.2011.03.002.

Morrison, G.S., 2013. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Aust. J. Forens. Sci.* 45, 173–197. http://dx.doi.org/10.1080/00450618.2012.733025.

Morrison, G.S., 2018. The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forens. Sci. Int.* 283, e1–e7. http://dx.doi.org/10.1016/j.forsciint.2017.12.024.

Morrison, G.S., Enzinger, E., 2019. Introduction to forensic voice comparison. In: Katz, W.F, Assmann, P.F. (Eds.), The Routledge Handbook of Phonetics. Routledge, Abingdon, UK, pp. 599–634. https://doi.org/10.4324/9780429056253-22.

Morrison, G.S., Enzinger, E., Zhang, C., 2016. Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015), the importance of distinguishing information from evidence/observations when formulating propositions. *Sci. Justice* 56, 492–497. http://dx.doi.org/10.1016/j.scijus.2016.07.002.

Morrison, G.S., Enzinger, E., Zhang, C., 2018. Forensic speech science. In: Freckelton, I., Selby, H. (Eds.), Expert Evidence. Thomson Reuters, Sydney, Australia Ch. 99.

Morrison G.S., Enzinger E., Ramos D., González-Rodríguez J., Lozano-Díez A. (in press). Statistical models in forensic voice comparison. In Banks D.L., Kafadar K., Kaye D.H., Tackett M. (eds.), *Handbook of Forensic Statistics*. Boca Raton, FL: CRC.

Morrison, G.S., Poh, N., 2018. Avoiding overstating the strength of forensic evidence: shrunk likelihood ratios/Bayes factors. *Sci. Justice* 58, 200–218. http://dx.doi.org/10.1016/j.scijus.2017.12.005.

Pigeon, S., Druyts, P., Verlinde, P., 2000. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digit. Signal Process.* 10, 237–248. http://dx.doi.org/10.1006/dspr.1999.0358.

Prince, S.J.D., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the IEEE 11th International Conference on Computer Vision, pp. 1–8. https://doi.org/10.1109/ICCV.2007.4409052.

Sizov, A., Lee, K.A., Kinnunen, T., 2014. Unifying probabilistic linear discriminant analysis variants in biometric authentication. In: Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M. (Eds.), Structural, Syntactic, and Statistical Pattern Recognition. Springer, Berlin, pp. 464–475. https://doi.org/10.1007/978-3-662-44415-3_47.